

Physics on autopilot: exploring the use of an AI assistant for independent problem-solving practice

Andrii V. Riabko¹, Tetiana A. Vakaliuk^{2,3,4,5}

¹Oleksandr Dovzhenko Hlukhiv National Pedagogical University, 24 Kyivska Str., Hlukhiv, 41400, Ukraine

²Zhytomyr Polytechnic State University, 103 Chudnivsyka Str., Zhytomyr, 10005, Ukraine

³Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine

⁴Kryvyi Rih State Pedagogical University, 54 Gagarin Ave., Kryvyi Rih, 50086, Ukraine

⁵Academy of Cognitive and Natural Sciences, 54 Gagarin Ave., Kryvyi Rih, 50086, Ukraine

Abstract. This study investigates the efficacy of large language model (LLM)-powered chatbots in guiding physics problem-solving, examining whether they can effectively supplement teacher-led learning. A customised chatbot was developed leveraging ChatGPT to provide step-by-step assistance through a structured problem-solving algorithm. Its impact was evaluated via an experimental study with 12th-grade physics students ($N = 24$) randomly assigned to a teacher-guided or chatbot-guided group for problem-solving practice. A Mann-Whitney U test revealed no significant differences in problem-solving competency between conditions. Qualitative analysis of conversational logs indicates the chatbot successfully emulated key teacher scaffolding behaviours. Our findings suggest AI tutors can deliver personalised, interactive support akin to human teachers, offering viable supplements to augment physics learning. Further research should explore optimising LLM training, human-chatbot balances, and impacts across diverse educational settings.

Keywords: large language models, physics education, chatbots, artificial intelligence, problem-solving

1. Introduction

Physics is a conceptually and mathematically rigorous scientific discipline, often posing significant challenges for students as they grapple with abstract theories and complex quantitative problems [5]. A vital competency is the ability to systematically unpack and solve the multifaceted physics word problems that exemplify theoretical principles [6]. However, many students need help to develop structured, deliberate problem-solving schemas instead of relying on superficial plug-and-chug methods that impede deep conceptual understanding [25].

Traditionally, physics instructors play a crucial role in scaffolding the development of structured problem-solving skills via personalised guidance and feedback [16]. However, resource constraints often limit individual teacher-student interactions, especially for large classes or distance learners. Could artificial intelligence (AI) tutors help bridge this gap by providing on-demand, interactive problem-solving support?

Recent advances in natural language processing have enabled the creation of AI systems with emergent linguistic capabilities and human-like dialogue skills [4]. Large language models

✉ ryabko@meta.ua (A. V. Riabko); tetianavakaliuk@gmail.com (T. A. Vakaliuk)

🌐 <http://acnsi.org/vakaliuk/> (T. A. Vakaliuk)

🆔 0000-0001-7728-6498 (A. V. Riabko); 0000-0001-6825-4697 (T. A. Vakaliuk)



© Copyright for this paper by its authors, published by Academy of Cognitive and Natural Sciences (ACNS). This is an Open Access article distributed under the terms of the Creative Commons License Attribution 4.0 International (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

(LLMs) like ChatGPT that can parse contextual information and generate responses mimicking natural conversation [9] are of particular interest. This raises the prospect of leveraging such AI to create personalised educational chatbots or virtual tutors, though their capabilities for complex tutoring remain relatively unexplored [23].

This study investigates the efficacy of using a customised, LLM-powered chatbot to guide students through physics problem-solving. It examines whether physics education chatbots can successfully supplement teacher-led scaffolding to enhance problem-solving skills. We hypothesise that a ChatGPT-based chatbot trained on a structured problem-solving approach will deliver individualised guidance on par with human teacher interventions.

2. Literature review

Problem-solving is a cornerstone of physics learning, requiring the methodical application of conceptual knowledge to quantitative situations [25]. However, students often rely on inflexible, plug-and-chug algorithms rather than adaptable, reflective strategies [6]. This superficial approach rarely imparts lasting conceptual gains [16].

Instructional interventions that model structured techniques and prompt metacognitive reflection are crucial for developing robust, flexible problem-solving schemas [5]. Traditionally, physics teachers play a crucial role in providing this scaffolding via personalised guidance and feedback [16]. However, resource constraints often limit these individualised interactions, highlighting the need for supplemental tutoring systems.

The educational landscape is poised for a profound transformation, fuelled by the burgeoning union of pedagogy and artificial intelligence (AI) [22]. The application of AI in education is experiencing a surge of interest, with momentum rapidly building across the field [18].

Intelligent tutoring systems (ITSs) have long aimed to fill this niche by leveraging artificial intelligence for customised instruction [17]. However, dialogue-based ITSs have proven difficult to implement due to challenges in natural language processing (NLP).

As indicated by Fuchs [7], NLP models offer tantalising prospects for personalised learning. Imagine curated lessons, tailored feedback, and readily available resources at every student's fingertips, wherever and whenever needed. However, this revolutionary potential comes with a responsibility to navigate potential pitfalls. To ensure optimal learning, universities must strike a delicate balance, leveraging NLP models as powerful supplements to, not replacements for, human interaction [7]. Recent advances in NLP have mitigated many obstacles through LLMs like ChatGPT, capable of remarkably human-like conversation [12].

LLMs like ChatGPT, GPT-3, and GPT-4 are neural networks trained on massive text datasets to generate coherent, context-responsive language. Their emergent reasoning and communication capabilities have fueled a growing interest in AI chatbots and virtual assistants for education [3]. A growing body of research on using ChatGPT in the higher education sector [2, 18, 20]. For instance, ChatGPT shows promising skills in answering science questions, explaining concepts, and solving problems, though its reliability needs to be better [11].

Jinchuñá Huallpa et al. [10] explore the ethical landscape of integrating ChatGPT into higher education, specifically delving into its implications for Latin American institutions. Their study highlights the nuanced perceptions of ChatGPT by stakeholders. While its accessibility and

potential for personalised learning are acknowledged, concerns regarding societal impact, data privacy, and ethical implementation remain prominent.

The swift and dynamic evolution of technologies like ChatGPT has significantly impacted global education systems. Ausat et al. [1] conclude that, within learning settings using ChatGPT, it is crucial to remember that technology exists as a tool, not a replacement for the irreplaceable role of the educator. As such, effective and appropriate technology integration into learning, alongside its continued development, is paramount.

Markel et al. [14] introduce GPTeach, an interactive chat-based tool simulating student interactions to train novice teachers. Two studies assessed its efficacy: a think-aloud study and an A/B test against a baseline. As teaching assistants during simulated office hours with GPT-powered students, participants benefited from GPTeach in several ways. First, it provided a risk-free environment for practice, shielding real students from potentially imperfect responses. Second, it facilitated iterative learning, allowing teachers to refine their responses within and across sessions. These findings suggest that GPTeach can be a valuable tool for enhancing teacher development.

While emerging studies have evaluated ChatGPT's standalone capabilities, few have examined its integration into larger pedagogical systems or experimental designs. Our study addresses this gap by implementing a ChatGPT-based chatbot for physics problem-solving within an experimental intervention, assessing its capacity to emulate and supplement teacher scaffolding behaviours. This provides empirical insight into its viability as an AI teaching aid.

3. Chatbot design

The current generation of virtual assistants can be broadly classified into task-oriented and conversational. While the former is designed to help users with specific tasks such as booking flights or ordering food, the latter aims to engage users in a conversation. However, to achieve the desired level of user engagement, conversational virtual assistants need to master the art of natural conversation. This requires them to maintain coherence and consistency throughout the conversation, much like a good friend. Achieving this balance is a challenging task. Traditionally, chatbots have relied on a three-step process: understanding the user's input, identifying the direction of the conversation, and selecting the best response from a pre-defined list. While this approach ensures that the chatbot does not make any errors, it can also make the conversation feel robotic and predictable.

To overcome this limitation, generative chatbots have been developed to respond to users word by word, similar to a seasoned translator. This approach promises a more natural flow of conversation but can also be unpredictable. We propose using dialogue ranking to find the sweet spot between predictability and naturalness. This involves curating a library of potential responses, each ranked based on how well it fits the current conversation. This approach ensures that the chatbot provides grammatically correct and relevant answers while maintaining flexibility to keep the conversation interesting.

In our experiment, we explored the effectiveness of dialogue ranking. We selected the most natural and engaging response by analysing the context of the user's input and comparing it to our pre-written responses. This approach can be considered matching puzzle pieces to

create a satisfying conversation. Ultimately, the ideal chatbot should master both structure and surprise. By prioritising coherence, consistency, and the right blend of pre-written and on-the-fly responses, we can create chatbots that perform tasks and leave users feeling like they have had a great conversation [15].

Imagine a chatbot that does not rely on a pre-written script but instead crafts responses on the spot, just like you do in a conversation. That is the idea behind generative chatbots. They are trained to predict each word in their responses, similar to how machine translation systems work [24].

As for the performance of generative models, they are somewhat unpredictable in their performance for use in commercial products [19]. Therefore, ranking models that choose from a pre-defined pool of answers are the most popular. For single-turn or multi-turn conversation pairs, vector representations of one dimension (encoder-encoder) are constructed from the data set. Then, the possible answers are ranked according to the values of some relevant function between the vectors (often a scalar product or cosine distance). This approach, which has gained popularity in information retrieval tasks, has subsequently been adapted in many works to create dialogue systems [21].

The world of chatbots has two main paths: the open-ended improvisation of generating models and the curated confidence of ranking models. While generating models can craft unique responses on the fly, their performance can be a wild card in commercial products [8]. That is why ranking models, which choose from a pre-defined pool of answers, reign supreme.

Ranking models are like treasure hunters, diving into a treasure chest of potential replies and analysing your message to find the best match. Using fancy algorithms called “vector representations”, they measure how relevant each answer is to your words. Think of it like checking for keywords or hidden connections. The best match wins, ensuring a grammatically correct and appropriate response every time. This approach is not just reliable; it is also familiar. Ranking models borrow techniques from information retrieval, where finding the perfect document for your search query is critical. By adapting these proven methods to the world of chatbots, we create smooth and familiar conversations without risking any awkward surprises.

Of course, both approaches have their strengths and weaknesses. However, ranking models offer a treasure trove of benefits for commercial products seeking consistent quality and user trust. They provide confidence for both users and developers, ensuring every interaction is a delightful dance of understanding, not a risky improvisation. Dialogue ranking agents select answers from among pre-prepared responses. Hence, an essential advantage of this approach is the ability to limit the output of grammatically incorrect and unacceptable answers that may be present in the training dataset. In our experiment, we chose a chatbot ranking implementation [13].

ChatGPT, a generative chatbot, has been proposed to teach students how to solve physics problems. The chatbot can be used in different ways, such as providing solutions to problems, suggesting possible solutions, explaining concepts or formulas, or providing additional resources. This approach can be helpful for students who have difficulty solving problems or want to improve their problem-solving skills.

ChatGPT has several advantages for teaching students to solve physics problems. Firstly, it is fast and can quickly solve problems, allowing students to get help quickly. Secondly, it can

be used on any device with internet access, making it accessible to many students. Thirdly, it can provide assistance tailored to the student's needs, making it a personalised learning tool. However, there are also limitations to be aware of.

Firstly, ChatGPT can only sometimes provide exact solutions to problems. It can only provide solutions to problems present in the dataset it was trained in. Secondly, ChatGPT may only sometimes be able to explain why it provides a particular solution.

In our research, we used ChatGPT as a basis for a chatbot that teaches the student how to solve a problem but only provides a ready answer at the intermediate stages of solving and calculation. Instead, it manages the student's work according to the algorithm for solving a physical problem developed and tested by the authors earlier.

Earlier, we considered the creation of a chatbot for teaching students how to solve physics problems, which worked based on a developed algorithm for solving a physical problem [21]. This study presents a novel physics problem-solving chatbot leveraging a structured, human-like approach. The system incorporates 11 key steps:

1. *Initial reading and clarification*: the chatbot carefully parses the problem statement, prompting the student for clarification on new terms or ambiguous expressions.
2. *Formalisation and unit conversion*: the problem statement is concisely documented, and all physical quantities are converted to the si system for consistency.
3. *Visualisation*: the system generates relevant diagrams, graphs, or figures to aid conceptual understanding.
4. *Physics analysis*: the chatbot guides the student in identifying the underlying physical phenomena and recalling relevant laws and formulas.
5. *Solution method selection*: the system collaborates with the student to determine the optimal solution method (analytical, synthetic, or mixed).
6. *Solution plan development*: a step-by-step plan is formulated to systematically solve the problem.
7. *Formula-based representation*: known and unknown quantities are systematically related using appropriate formulas.
8. *Equation solving*: the chatbot assists in solving equations or systems of equations to obtain the final formula.
9. *Numeric solution*: the desired value is accurately calculated based on the derived formula.
10. *Solution validation*: the results are critically analysed for consistency and reasonableness with the student.
11. *Alternative approaches*: the system encourages exploration of alternate solution paths to promote deeper understanding and flexibility (figure 1).

The structured approach of teaching physics problems using a chatbot mimics the expert thought process, aiming to solve problems and equip students with a comprehensive understanding of physics concepts and problem-solving strategies.

However, creating a chatbot for each specific task requires much effort. Traditionally, this has involved writing down possible options for dialogue with the student. However, using GPT-based chatbots can solve this problem quickly. Programming the dialogues will be done by artificial intelligence, which is enough to write detailed prompts.

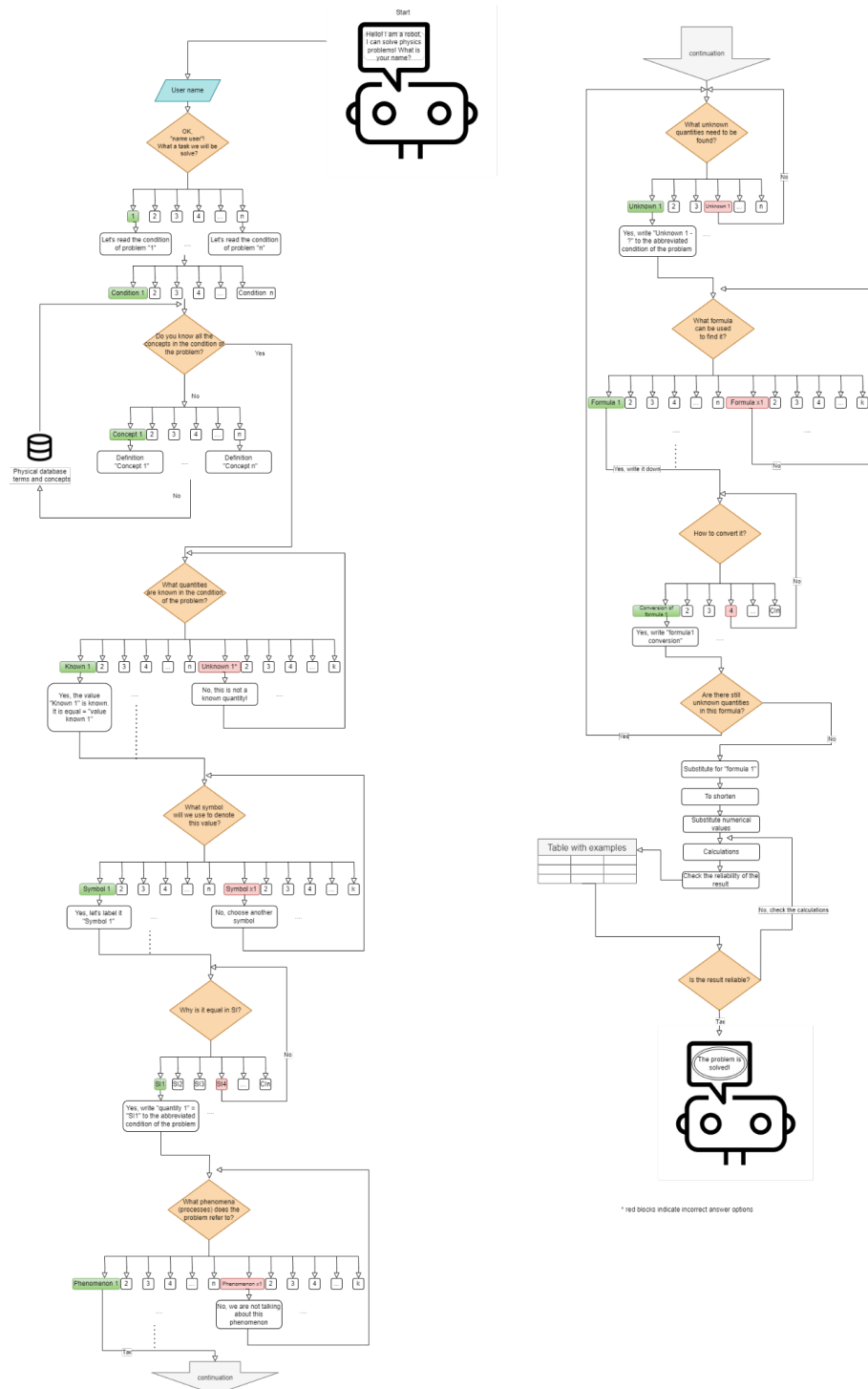


Figure 1: Flowchart of the chatbot algorithm for solving physics problems.

It is worth noting that ChatGPT is excellent at solving typical physics problems. This is the subject of a possible study, which could quickly lose relevance as new technologies based on artificial intelligence are being improved very quickly. Therefore, we will limit ourselves to some convincing examples.

It is proposed to solve the following problem:

- The velocity of liquid flow in a certain section of a horizontal pipe is $v_1 = 5$ cm/s. Find the flow velocity in the pipe's part with half the smaller diameter.

The junction is shown in the figure 2.

ChatGPT also performs intermediate actions in solving the problem, for example, transferring physical values to the SI system (figure 3).

For the sake of fairness, he does not solve all the problems of GPT correctly. However, in all cases, he defines the physical phenomenon referred to in the condition of the problem, correctly writes down the equations and formulas that describe this phenomenon, and performs the solution in the correct sequence.

Integrating GPT-3 or GPT-4 with a Telegram bot involves creating a backend server that communicates with the OpenAI API and handles the interactions with your bot. Below are general steps you can follow to connect GPT to your Telegram bot:

1. Create a Telegram bot:
 - Talk to the BotFather on Telegram to create a new bot and obtain the API token.
 - Note down the token, as you will need it to communicate with Telegram's Bot API.
2. Set up a backend server:
 - You need a server to handle Telegram and the OpenAI API communication.
 - This can be done using a serverless platform like AWS Lambda or a traditional server.
3. Develop a Telegram bot backend: write code (e.g., using Python with a library like python-telegram-bot) to handle incoming messages from Telegram and send requests to the OpenAI API.
4. Integrate OpenAI API:
 - Use the OpenAI API to interact with GPT-3 or GPT-4.
 - You will need the OpenAI API key, which you can obtain by signing up on the OpenAI platform.
5. Process Telegram messages: When a user sends a message to your Telegram bot, your backend should process the message, send it to the OpenAI API to generate a response, and then send the generated response back to the user via the Telegram API.
6. Handle conversational context: you may need to store and manage conversation history in your backend to maintain context in conversations.
7. Deploy and test:
 - Deploy your backend server to a hosting provider.
 - Test your Telegram bot to ensure it works as expected.

ChatGPT 3.5 ▾

**You**

Solve this physics problem:

The velocity of liquid flow in a certain section of a horizontal pipe is $v_1=5$ cm/s. Find the flow velocity in the part of the pipe that has: half the smaller diameter

**ChatGPT**

To solve this problem, we can use the principle of conservation of mass for an incompressible fluid. The mass flow rate (Q) remains constant along the pipe. The mass flow rate is given by the product of the cross-sectional area (A) and the flow velocity (v).

Mathematically, $Q = A \cdot v$.

The cross-sectional area of the pipe is related to its diameter (D) by the formula $A = \frac{\pi D^2}{4}$.

Now, let's consider the case where the diameter is halved:

a) Half the smaller diameter:

If the diameter is halved, the new diameter (D_2) is given by $D_2 = \frac{D}{2}$. Therefore, the new cross-sectional area (A_2) is $A_2 = \frac{\pi (\frac{D}{2})^2}{4} = \frac{\pi D^2}{16}$.

Now, let's find the new velocity (v_2) using the conservation of mass equation:

$$Q = A_1 \cdot v_1 = A_2 \cdot v_2$$

$$\frac{\pi D^2}{4} \cdot v_1 = \frac{\pi D^2}{16} \cdot v_2$$

$$v_2 = 4 \cdot v_1$$

So, the flow velocity in the part of the pipe with half the smaller diameter is 4×5 cm/s = 20 cm/s.



Figure 2: Solution of the problem with the ChatGPT 3.5.

Here is a simplified Python example using the python-telegram-bot library and the OpenAI API:

```
from telegram import Update
from telegram.ext import Updater, CommandHandler, MessageHandler
from telegram.ext import Filters, CallbackContext
import openai
```

ChatGPT 3.5 ▾

**ChatGPT**

To translate the physical quantities to the International System of Units (SI), we need to convert the given velocity from centimeters per second to meters per second.

Given velocity in the problem: $v_1 = 5 \text{ cm/s}$

Conversion:

$$1 \text{ cm} = 0.01 \text{ m}$$

So, the velocity in SI units ($v_{1\text{SI}}$) is given by:

$$v_{1\text{SI}} = 5 \text{ cm/s} \times 0.01 \text{ m/cm} = 0.05 \text{ m/s}$$

Now, when solving the problem in the SI system, the velocity (v_2a) in the part of the pipe with half the smaller diameter will also be in meters per second.



Figure 3: Conversion to the SI system with the ChatGPT 3.5.

```
import os

# Set your Telegram bot token
TELEGRAM_BOT_TOKEN = 'your_telegram_bot_token'
# Set your OpenAI API key
OPENAI_API_KEY = 'your_openai_api_key'
# Initialize OpenAI API
openai.api_key = OPENAI_API_KEY

# Define the command handler
def start(update: Update, context: CallbackContext) -> None:
    update.message.reply_text('Hello! I am your bot. Send me a message.')

# Define the message handler
def handle_message(update: Update, context: CallbackContext) -> None:
    # Get user input
    user_input = update.message.text
    # Send user input to OpenAI API
    response = openai.Completion.create(engine="text-davinci-003",
        prompt=user_input, max_tokens=50, temperature=0.7)
    # Extract the generated response
    bot_response = response['choices'][0]['text']
    # Send the bot response to the user
    update.message.reply_text(bot_response)
```

```
def main() -> None:
    # Set up the Telegram updater
    updater = Updater(token=TELEGRAM_BOT_TOKEN, use_context=True)
    dispatcher = updater.dispatcher
    # Add command and message handlers
    dispatcher.add_handler(CommandHandler("start", start))
    dispatcher.add_handler(MessageHandler(Filters.text &
                                          ~Filters.command, handle_message))

    # Start the bot
    updater.start_polling()
    # Run the bot until you send a signal to stop it
    updater.idle()

if __name__ == '__main__':
    main()
```

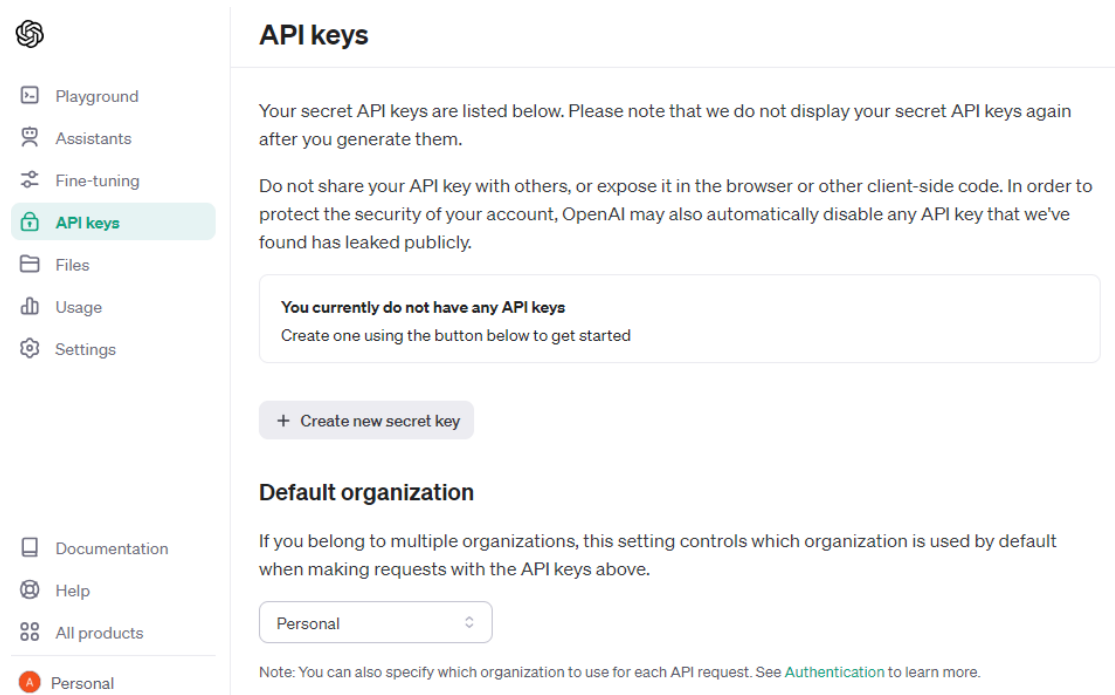


Figure 4: API key generation on <https://platform.openai.com/api-keys>.

Remember to replace 'your_telegram_bot_token' and 'your_openai_api_key' with your Telegram bot token and OpenAI API key.

Several AI models can be used to generate answers: ChatGPT (gpt-3.5-turbo), ChatGPT (gpt-3.5-turbo-16k), ChatGPT (gpt-3.5-turbo-16k-instruct), User fine-tuned model, User customised

model (Instruct), GPT-4.

Fine-tuning a chatbot's behaviour requires detailed instructions to govern its responses, set conversational tone, establish custom personas, and integrate brand information:

- Each potential interaction scenario should be defined with extensive context and conditional branching. Specify user personas, problem-solving stages, response prerequisites, sensitive topics, and relevant details to consider during answer generation. Encourage the model to generate multiple possible responses for comparison and selection. Guide it towards the most suitable output based on desired characteristics.
- Clearly illustrate your objectives using concrete examples. Specify desired response formats, ranking algorithms, sentiment classification targets, and question-answering styles. Provide example questions and anticipated responses for conversational tone guidance. Train the model with accurate and diverse data sets. Review example data for inconsistencies and potential biases that may influence response accuracy. Specify target languages and avoid numerical values whenever possible to maximise interpretability. If a specific persona is desired, describe their background, personality, communication style, and relevant life experiences. This enables the model to mimic their conversational traits and enhance user engagement.
- Thoroughly test and evaluate the trained model through user interaction simulations. Analyse conversational logs and update instructions to refine control, eliminate unwanted responses, and improve overall performance. Adopting these strategies allows you to create a highly responsive and personalised chatbot that fulfils its intended purpose.

GPT models can perform various tasks — from complex text analysis to generating an answer to an unlimited list of topics. To limit the chatbot's response area, set the tone of the conversation, personalise your bot to a specific character or person, and add information about your company. You need to add instructions for the bot.

When creating a tip, consider the following recommendations:

- Add the maximum context and conditions to the answer in each scenario. List all the prerequisites for interaction with the bot: indicate which users and at what stage of solving their problems they will contact, which details should be included in the answers, and which topics should be avoided.
- Let the task model generate several results to compare and specify the most suitable one.
- Make what you want clear with examples. For example, if you need the model to alphabetically rank a list of items or classify a paragraph by sentiment, list examples of requests, the expected format of the result, or what effect you want to achieve. If you need the bot to answer questions in a certain way, give an example of a question and an answer.
- Provide high-quality and accurate data. Check your examples – the model is usually intelligent enough to recognise basic spelling mistakes, but it can also assume that it was done on purpose, which can affect the answer. If you need the model to respond in a specific language, specify that language directly. It is also recommended to use words instead of numbers. Remember that AI understands all instructions literally.

The screenshot shows the 'OpenAI (ChatGPT)' configuration window. At the top, there's a close button (X). Below it, the 'API Secret key' section has a dropdown menu set to 'Use the token from the account settings' and a status bar indicating 'Free requests available: 30' with a link to 'Settings'. The 'CHATGPT: AI TEXT RESPONSE GENERATION' section has a green toggle switch turned on, with a note: 'After enabling the integration, the AI bot from OpenAI will respond instead of the Standard Reply flow.' Below this is the 'Model' section with a dropdown menu set to 'ChatGPT (gpt-3.5-turbo)' and a note: 'GPT-3.5 is the fastest and cheapest text model, GPT-4 is the most trained but more expensive.' with a link to 'Learn more'. The 'Bot instructions (Prompt)' section has a text area containing: 'The bot responds only in English. You are a physics teacher. Your task is to teach the student (bot user) to solve physics problems. Do not give the solution to the problem right away. Ask the user to wait for the solution in the following way:'. To the right of this text area is a link to 'Documentation'. At the bottom, there are two input fields: 'Maximum number of tokens in response' set to '32' with a note '~ 24 words' and 'Temperature' set to '1' with a note 'Between 0 and 2.'.

Figure 5: Creating instructions for the bot (prompts).

- Personalise the model. So that the model can enter the image of a specific person or character, describe what he does, what qualities he possesses, in what tone or with what lexicon he will communicate, as well as other aspects of the virtual assistant's life.
- Test the result and update the instructions. After setting up the prompts, test the result, view dialogues with users, and, if necessary, adjust the bot's instructions by adding or removing details. Train the model until you achieve the desired results.

Temperature is a parameter affecting the abstractness of answers. For example, if you ask the same question, the initial data will be different according to the selected temperature – abstract or more precise. Higher values closer to 2, for example, 1.3, will make the answers more random. Lower values closer to 0, for example, 0.2, will make them more comprehensive but the same in meaning.

The traditional approach is to use prompt engineering. In our case, the system instructions of the bot, which controls the operation of the model during the conversation, will look like this:

Bot only answers in English. You are a physics teacher. Your task is to teach the student (bot user) to solve physics problems. Wait to give the solution to the problem. Ask the

user questions one at a time in the following sequence.

Write to me about the condition of the problem in full.

What are the physical quantities of the species in the problem?

Expressions of known quantities in the SI system.

If necessary, make a drawing or diagram. What to depict on the diagram?

What physical quantities should be found in the problem?

What physical phenomena or processes are mentioned in the condition of the problem?

What formulas describe this phenomenon?

Which of the formulas can be used to solve the problem?

Make a system of equations to find an unknown quantity.

Express the unknown quantity from the system of equations.

Check the dimensionality of the calculation formula.

Substitute numerical values for the calculation formula.

Calculate the result.

Analyse the reliability of the obtained results and write down the answer.

The instructions provided are indicative and may vary depending on the physics section. Paying particular attention to tasks requiring a written description of a drawing or diagram is recommended. However, the limited context of GPT-3, which can only accept 4096 tokens, poses a challenge. Moreover, this approach is tedious as it requires constant manual content addition.

Suppose ChatGPT needs to transfer hundreds of PDF documents using physics techniques. In that case, the problem of a paywall, which restricts access to the content of a web page until payment is made, will soon arise. One way to overcome these restrictions is through a Python library called LlamaIndex [26].

Extending ChatGPT with LlamaIndex enables the configuration of ChatGPT chatbots to work with data from documents. This opens up new possibilities for creating chatbots that can communicate in a conversational style and perform tasks that require understanding context. LlamaIndex positions itself as a crucial tool for unlocking the full potential of large language models by bridging diverse data sources and facilitating contextual learning. It achieves this through the following key mechanisms: Establishing seamless connections with various data sources (APIs, PDFs, documents, SQL, etc.) using a comprehensive set of data

connectors and indexing both structured and unstructured data, ensuring accessibility for LLM processing, constructing indexes that optimise context awareness for LLMs—mitigating common challenges associated with LLM training, such as repetitive patterns and pain points, retaining context in an easily retrievable format, enabling rapid insertion as needed. Addressing tokens is constrained within LLMs (4096 for GPT-3 Davinci, 8000 for GPT-4) by significantly augmenting contextual information. LlamaIndex has the potential to alleviate text fragmentation issues and enable seamless user-index interactions. It simplifies extracting pertinent content from documents, facilitating prompt generation and response optimisation. LlamaIndex holds significant potential to enhance the contextual learning capabilities of LLMs, ultimately leading to more comprehensive, accurate, and contextually relevant responses. Its ability to integrate diverse data sources and optimise context management makes it a valuable asset in advancing LLM-powered applications [26].

4. Results

The detailed experimental verification of the described technique consisted of the following. In the experimental group of students (10th grade, 12 students), learning how to solve physics problems was taught using the chatbot developed based on ChatGPT. Students could use it to solve problems independently at school or home. Teaching was carried out in the control group (another 10th grade, 12 students) according to the traditional method. The teacher explained to students individually or in groups about solving problems. At the end of the experiment (1 school semester), a written test was conducted, which contained five problems from the sections of physics studied in this semester. Participation was voluntary with informed parental consent. Evaluation of works was carried out by a group of experts (teachers, methodologists, teachers of physics methods from the pedagogical university) on a 12-point scale (table 1).

Initial ratings were rank-ordered, and the Mann-Whitney U test was used to compare ranks for $n = 12$ participants in condition A and $n = 12$ participants in condition B .

Because the sample sizes are small and they suspect that the sample distribution is not normal, we decided to perform the Mann-Whitney U test to determine whether there is a statistically significant difference between the test scores of the control and experimental groups.

We can formulate hypotheses based on a consistent and systematic difference between the two methods of influence being compared (traditional and experimental).

1. Null hypothesis (H_0): absence of a significant difference between the two conditions. This implies no tendency for ranks in one condition to be systematically higher (or lower) than ranks in the other.
2. Alternative hypothesis (H_1): a significant difference between the two conditions. This suggests that the scores in one condition are systematically higher (or lower) than those in the other.

After the testing, they ranked the grades and wrote separate ranks for samples A (control) and B (experimental). Then we find $\sum R_A$, the sum of students' ranks in sample A , and $\sum R_B$ for sample B . For sample A , the sum of ranks is $\sum R_A = 152$. For sample B , the sum of ranks is $\sum R_B = 148$ (table 1).

Table 1

Results of final testing.

№	Scores in the control group	Scores in the experimental group	Control group (ranks)	Experimental group (ranks)
1	12	10	24	20.5
2	9	10	14	20.5
3	9	8	14	8
4	10	9	20.5	14
5	6	9	2	14
6	7	8	5	8
7	9	11	14	23
8	9	6	14	2
9	6	7	2	5
10	8	9	8	14
11	9	7	14	5
12	10	9	20.5	14
			$\sum R_A = 152$	$\sum R_B = 148$

Then, the values of U for samples A and B were calculated:

$$U_A = n_A \cdot n_B + \frac{n_A(n_A + 1)}{2} - \sum R_A. \quad (1)$$

$$U_B = n_A \cdot n_B + \frac{n_B(n_B + 1)}{2} - \sum R_B. \quad (2)$$

For sample A , $U_A = 70$. For sample B , $U_B = 74$. The Mann-Whitney U test is 70.

The critical value of the Mann-Whitney U test for a given number of compared groups is 37.

$70 > 37$; therefore, the differences in the level of the trait in the compared groups are not statistically significant ($p > 0.05$).

Thus, we cannot support the H_0 hypothesis because the data do not provide sufficient evidence to establish an actual difference between the two minds.

Experimental verification of the developed method of applying artificial intelligence technology in a chatbot for learning how to solve problems in physics has proven its effectiveness. It is recommended for implementation in the educational process. A chatbot based on artificial intelligence can perform part of the teacher's functions in teaching students to solve physics problems.

5. Discussion

This study provides experimental evidence that LLM chatbots can deliver practical physics problem-solving support, rivalling traditional teacher scaffolding. The chatbot's success in replicating personalised, responsive tutoring interactions supports AI's viability for customised learning at scale, complementing resource-limited human teachers.

However, limitations exist regarding LLM reliability, training scalability, and generalizability. Chatbots are only as effective as their training data, like any AI system. Thorough curation and ongoing tweaking of knowledge bases and dialogue models are essential to ensure accurate, high-quality responses. Striking an optimal balance between pre-scripted algorithms and generative capabilities remains an open challenge.

An additional study of the chatbot features based on ChatGPT regarding the number of incorrect recommendations for solving problems is required. It requires the development of methods for training a chatbot on selected texts from the physics methodology and creating a database of texts in the required format, especially concerning the presentation of formulas in the texts.

The limited sample size necessitated using the Mann-Whitney U test, a non-parametric alternative that does not rely on assumptions of normality or homogeneity of variances. While this offers flexibility, some caveats deserve attention. One fundamental weakness is its sensitivity to tied ranks, meaning identical values within or across groups. The presence of numerous ties can compromise the test's accuracy. Additionally, the Mann-Whitney U test cannot directly quantify the magnitude of the observed difference between groups (effect size). To address this, complementary methods like Spearman's rank correlation coefficient or Wilcoxon's effect size offer valuable insights, though their application is often more appropriate for larger sample sizes.

6. Conclusions

This study investigates the potential of ChatGPT-powered chatbots to revolutionise physics education by partially supplanting traditional teacher-led learning in problem-solving. Our findings demonstrate that chatbots can effectively mimic essential teacher functions, guiding students through problem-solving algorithms and offering personalised feedback.

We developed a customised chatbot leveraging the ChatGPT API to provide interactive support during physics problem-solving. The system was designed to replicate key teacher scaffolding behaviours identified in the literature [5, 16]:

- Prompting conceptual analysis
- Guiding methodical execution
- Explaining rationales
- Tracing step-by-step logic
- Providing feedback on student responses
- Encouraging reflection and clarity

The chatbot was trained on a corpus of physics educational materials and sample dialogues demonstrating ideal tutoring approaches to implement these behaviours. Its responses were further fine-tuned via prompt engineering to align with the following problem-solving algorithm:

1. Parse problem statement
2. Identify known and target variables
3. Apply conceptual frameworks

4. Select appropriate equations
5. Mathematically represent relationships
6. Derive target quantity
7. Compute numerical solution
8. Assess logical validity

During interactions, the chatbot verbalises each step, provides tailored hints and feedback, answers clarifying questions, and adapts follow-up prompts based on student replies. This structured technique actively guides students through deliberate, reflective problem-solving while exploring their grasp of underlying concepts and strategies.

An experimental study was conducted with 10th-grade physics students to compare chatbots' effectiveness against traditional teacher-led problem-solving instruction. Two groups of students tackled physics problems: the teacher-led group received traditional explanations and guidance from a teacher, and the chatbot group solved problems independently, assisted by a custom-built chatbot based on a generalised problem-solving sequence. The chatbot monitored solution correctness, posed leading questions, and tracked student responses, replicating a teacher's role in learning. The Mann-Whitney U test revealed no significant difference in problem-solving skills between the groups. This suggests that the chatbot successfully emulated the effectiveness of teacher consultations and explanations.

Qualitative analysis of chatbot dialogue logs indicates the system successfully provided step-by-step guidance, feedback, and explanations akin to human tutoring behaviours. Instances of flawed reasoning were rare and successfully self-corrected through continued dialogue.

Our findings pave the way for integrating chatbots like ChatGPT into physics education. These AI-powered tutors can supplement traditional learning by providing personalised support, freeing teachers for more individualised attention. While not a complete replacement for human interaction, chatbots offer exciting possibilities for enhancing engagement and boosting problem-solving skills in future physics classrooms.

Key directions include expanding training datasets, integrating educator and student feedback loops, and conducting large-scale trials across diverse educational settings and populations. As LLMs evolve rapidly, researchers must run alongside rigorously evaluating implications for learning and illuminating best practices to maximise benefits while minimising risks. Future studies should explore the long-term impact of chatbot integration on student learning outcomes and optimal methods for incorporating these tools into diverse educational settings.

Acknowledgments

This research was supported by the use of AI-assisting tools (LLMs GPT 3.5 and 4.0), specifically in generating outlines, suggesting relevant sections, and identifying potential gaps in the argument. We carefully reviewed and revised the generated content, ensuring it aligned with our research and ethical standards while incorporating relevant, established concepts. We acknowledge that all original ideas and text remain our own and express gratitude to OpenAI and Grammarly for enabling a responsible and creative use of artificial intelligence in this research.

References

- [1] Ausat, A.M.A., Massang, B., Efendi, M., Nofirman, N. and Riady, Y., 2023. Can Chat GPT Replace the Role of the Teacher in the Classroom: A Fundamental Analysis. *Journal on Education*, 5(4), pp.16100–16106. Available from: <https://jonedu.org/index.php/joe/article/view/2745>.
- [2] Barbas, M.P., Vieira, A.T. and Branco, P.D., 2024. The Importance of Chat GPT Training for Higher Education: Case Study. In: N. Martins and D. Brandão, eds. *Advances in Design and Digital Communication IV*. Cham: Springer Nature Switzerland, *Springer Series in Design and Innovation*, vol. 35, pp.695–705. Available from: https://doi.org/10.1007/978-3-031-47281-7_57.
- [3] Bitzenbauer, P., 2023. ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, 15(3), p.ep430. Available from: <https://doi.org/10.30935/cedtech/13176>.
- [4] Budzianowski, P., Ultes, S., Su, P.H., Mrkšić, N., Wen, T.H., Casanueva, I., Rojas-Barahona, L.M. and Gašić, M., 2017. Sub-domain Modelling for Dialogue Management with Hierarchical Reinforcement Learning. In: K. Jokinen, M. Stede, D. DeVault and A. Louis, eds. *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Saarbrücken, Germany: Association for Computational Linguistics, pp.86–92. Available from: <https://doi.org/10.18653/v1/W17-5512>.
- [5] Chi, M.T.H., 2005. Commonsense Conceptions of Emergent Processes: Why Some Misconceptions Are Robust. *The Journal of the Learning Sciences*, 14(2), pp.161–199. Available from: <https://www.public.asu.edu/~mtchi/papers/EmergJLSpdf.pdf>.
- [6] Docktor, J.L., Strand, N.E., Mestre, J.P. and Ross, B.H., 2015. Conceptual problem solving in high school physics. *Phys. Rev. ST Phys. Educ. Res.*, 11, p.020106. Available from: <https://doi.org/10.1103/PhysRevSTPER.11.020106>.
- [7] Fuchs, K., 2023. Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse? *Frontiers in Education*, 8. Available from: <https://doi.org/10.3389/feduc.2023.1166682>.
- [8] Gao, J., Galley, M. and Li, L., 2018. Neural Approaches to Conversational AI. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, SIGIR '18, p.1371–1374. Available from: <https://doi.org/10.1145/3209978.3210183>.
- [9] Hamaniuk, V.A., 2021. The potential of Large Language Models in language education. *Educational Dimension*, 5, p.208–210. Available from: <https://doi.org/10.31812/ed.650>.
- [10] Jinchuña Huallpa, J., Flores Arocutipa, J.P., Panduro, W.D., Chauca Huete, L., Flores Limo, F.A., Espinoza Herrera, E., Alba Callacna, R.A., Ariza Flores, V.A., Medina Romero, M. Ángel, Merino Quispe, I. and Hernández Hernández, F.A., 2023. Exploring the ethical considerations of using Chat GPT in university education. *Periodicals of Engineering and Natural Sciences*, 11(4), pp.105–115. Available from: <http://pen.ius.edu.ba/index.php/pen/article/view/3770>.
- [11] Kieser, F., Wulff, P., Kuhn, J. and Küchemann, S., 2023. Educational data augmentation in physics education research using ChatGPT. *Phys. Rev. Phys. Educ. Res.*, 19, p.020150. Available from: <https://doi.org/10.1103/PhysRevPhysEducRes.19.020150>.

- [12] Liang, Y., Zou, D., Xie, H. and Wang, F.L., 2023. Exploring the potential of using ChatGPT in physics education. *Smart Learning Environments*, 10(1), p.52. Available from: <https://doi.org/10.1186/s40561-023-00273-7>.
- [13] Logacheva, V., Malykh, V., Litinsky, A. and Burtsev, M., 2020. ConvAI2 Dataset of Non-goal-Oriented Human-to-Bot Dialogues. In: S. Escalera and R. Herbrich, eds. *The NeurIPS '18 Competition*. Cham: Springer International Publishing, The Springer Series on Challenges in Machine Learning, pp.277–294. Available from: https://doi.org/10.1007/978-3-030-29135-8_11.
- [14] Markel, J.M., Opferman, S.G., Landay, J.A. and Piech, C., 2023. GPTeach: Interactive TA Training with GPT-Based Students. *Proceedings of the Tenth ACM Conference on Learning @ Scale*. New York, NY, USA: Association for Computing Machinery, L@S '23, p.226–236. Available from: <https://doi.org/10.1145/3573051.3593393>.
- [15] Masche, J. and Le, N.T., 2018. A Review of Technologies for Conversational Systems. In: N.T. Le, T. van Do, N.T. Nguyen and H.A.L. Thi, eds. *Advanced Computational Methods for Knowledge Engineering*. Cham: Springer International Publishing, *Advances in Intelligent Systems and Computing*, vol. 629, pp.212–225. Available from: https://doi.org/10.1007/978-3-319-61911-8_19.
- [16] Mestre, J.P., 2002. Probing adults' conceptual understanding and transfer of learning via problem posing. *Journal of Applied Developmental Psychology*, 23(1), p.9–50. Available from: [https://doi.org/10.1016/s0193-3973\(01\)00101-0](https://doi.org/10.1016/s0193-3973(01)00101-0).
- [17] Nkambou, R., Bourdeau, J. and Mizoguchi, R., eds, 2010. *Advances in Intelligent Tutoring Systems, Studies in Computational Intelligence*, vol. 308. Berlin, Heidelberg: Springer Berlin Heidelberg. Available from: https://doi.org/10.1007/978-3-642-14363-2_1.
- [18] Oranga, J., 2023. Benefits of artificial intelligence (ChatGPT) in education and learning: Is Chat GPT helpful? *International Review of Practical Innovation, Technology and Green Energy (IRPITAGE)*, 3(3), p.46–50. Available from: <https://radjapublika.com/index.php/IRPITAGE/article/view/1250>.
- [19] Prananta, A.W., Megahati S, R.R.P., Susanto, N. and Raule, J.H., 2023. Transforming Education and Learning through Chat GPT: A Systematic Literature Review. *Jurnal Penelitian Pendidikan IPA*, 9(11), p.1031–1037. Available from: <https://doi.org/10.29303/jppipa.v9i11.5468>.
- [20] Rakap, S., 2023. Chatting with GPT: Enhancing Individualized Education Program Goal Development for Novice Special Education Teachers. *Journal of Special Education Technology*, p.01626434231211295. Available from: <https://doi.org/10.1177/01626434231211295>.
- [21] Riabko, A.V., Vakaliuk, T.A., Zaika, O.V., Kukharchuk, R.P. and Kontsedailo, V.V., 2023. Chatbot algorithm for solving physics problems. In: T.A. Vakaliuk, V.V. Osadchy and O.P. Pinchuk, eds. *Proceedings of the 2nd Workshop on Digital Transformation of Education (DigiTransfEd 2023) co-located with 18th International Conference on ICT in Education, Research and Industrial Applications (ICTERI 2023), Ivano-Frankivsk, Ukraine, September 18-22, 2023*. CEUR-WS.org, *CEUR Workshop Proceedings*, vol. 3553, pp.75–92. Available from: <https://ceur-ws.org/Vol-3553/paper5.pdf>.
- [22] Semerikov, S.O., Striuk, A.M. and Shalatska, H.M., 2021. AI-assisted language education: critical review. *Educational Dimension*, 4, p.1–7. Available from: <https://doi.org/10.31812/ed.623>.

- [23] Tong, D., Tao, Y., Zhang, K., Dong, X., Hu, Y., Pan, S. and Liu, Q., 2023. Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. *Asia Pacific Education Review*. Available from: <https://doi.org/10.1007/s12564-023-09913-6>.
- [24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is All you Need. In: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan and R. Garnett, eds. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Curran Associates, Inc., vol. 30, pp.5998–6008. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [25] Walsh, L.N., Howard, R.G. and Bowe, B., 2007. Phenomenographic study of students' problem solving approaches in physics. *Phys. Rev. ST Phys. Educ. Res.*, 3, p.020108. Available from: <https://doi.org/10.1103/PhysRevSTPER.3.020108>.
- [26] Zirnstein, B., 2023. *Extended context for InstructGPT with LlamaIndex*. (Report). Available from: <https://www.researchgate.net/profile/Bruno-Zirnstein/publication/371911146>.